Factorizing Probabilistic Graphical Models Using Co-occurrence Rate

Zhemin Zhu
Department of Computer Science , Technische Universität Darmstadt
Hochschulstr.10, 64289, Darmstadt, Germany
{zhuzhemin@gmail.com}

Abstract

Factorization is of fundamental importance in the area of Probabilistic Graphical Models (PGMs). In this paper, we theoretically develop a novel mathematical concept, **C**o-occurrence **R**ate (CR), for factorizing PGMs. CR has three obvious advantages: (1) CR provides a unified mathematical foundation for factorizing different types of PGMs. We show that Bayesian Network Factorization (BN-F), Conditional Random Field Factorization (CRF-F), Markov Random Field Factorization (MRF-F) and Refined Markov Random Field Factorization (RMRF-F) are all special cases of CR Factorization (CR-F); (2) CR has simple probability definition and clear intuitive interpretation. CR-F tells not only the scopes of the factors, but also the exact probability functions of these factors; (3) CR connects probability factorization and graph operations perfectly. The factorization process of CR-F can be visualized as applying a sequence of graph operations including partition, merge, duplicate and condition to a PGM graph. We further obtain an important result: by CR-F, on TCG graphs the scopes of factors can be exactly over maximal cliques without any default configuration. This improves the results of (R)MRF-F which need default configurations, and also indicates that (R)MRF-F, as special cases of CR-F, can not always achieve the optimal results of CR-F.

1 Introduction

Independence is a very important type of experience that can be used to simplify PMs. PGMs are compact formalizations of independence relations among random variables which use different types of graphs as their representations. The fundamental problem in the area of PGMs is to factorize high dimensional joint probabilities into small factors based on the independence relations among random variables. Learning and inference algorithms are based on the results of factorization.

Bayesian networks (BNs) are directed acyclic graphs. The conditional independence of BNs can be judged by d-separation criteria (Pearl, 1986). BN-F is based on the mathematical concept of conditional probability:

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i, G)),$$

where $Pa(x_i, G)$ are all the parents of the node x_i in the BN graph G.

Markov networks (MNs) are undirected graphs which can contain cycles. According to the Markov property, a set of nodes are independent with non-adjacent nodes conditioned by their immediate neighbours which are called Markov Blanket (MB). MRF-F is based on the **H**ammersley-**C**lifford Theorem (Clifford, 1990) which tells a joint probability over a MN can always be written as a product of functions over all maximal cliques:

$$P(x_1, x_2, ..., x_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(mc_i),$$

where $\{mc_1, mc_2, ..., mc_m\}$ are all the maximal cliques; $\{\phi_1, \phi_2, ..., \phi_m\}$ are potential functions over maximal cliques; and Z is the partition function for normalization.

The HC Theorem can be proved in a constructive way (Cheung, 2008) by defining a candidate potential function as:

$$f_i(c_i) = \prod_{s \in \mathcal{P}(c_i)} P(X_s = x_s, X_{G \setminus s} = 0)^{-1^{|c_i| - |s|}}, \quad (1)$$

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{l} f_i(c_i), \quad (2)$$

where $\{c_1, c_2, ..., c_l\}$ are all cliques in G including the empty clique $\emptyset$; $\mathcal{P}(c_i)$ is the power set of c_i including

the boundary cases $\emptyset$ and c_i; $|*|$ is the number of nodes in $*$; and $P(X_s = x_s, X_{G\setminus s} = 0)$ is the joint probability with X_s set to the corresponding values x_s and the remainder of the graph $X_{G\setminus s}$ set to default configuration values denoted as 0. If we group the cliques into maximal cliques, then the potential functions over maximal cliques are:

$$\phi_i(mc_i) = \prod_{c_j \in \mathcal{P}(mc_i)} f_j(c_j)$$
$$= \prod_{c_j \in \mathcal{P}(mc_i)} \prod_{s \in \mathcal{P}(c_j)} P(X_s = x_s, X_{G\setminus s} = 0)^{-1^{|c_j|-|s|}}$$

If we replace the potential function over cliques in Eqn.(2) with Eqn.(1) and apply the Markov property, a Refined MRF Factorization (RMRF-F) can be obtained which can be represented as a factor graph (Abbeel *et al.*, 2005):

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{l} \prod_{s \in \mathcal{P}(c_i)} P(X_s = x_s, X_{G\setminus s} = 0)^{-1^{|c_i|-|s|}}$$
$$= \prod_{i=1}^{l} \prod_{s \in \mathcal{P}(c_i)} [P(X_s = x_s, X_{c_i\setminus s} = 0 | X_{G\setminus c_i} = 0)$$
$$P(X_{G\setminus c_i} = 0)]^{-1^{|c_i|-|s|}}$$
$$= \prod_{i=1}^{l} \prod_{s \in \mathcal{P}(c_i)} P(X_s = x_s, X_{c_i\setminus s} = 0 | X_{MB(c_i)} = 0)^{-1^{|c_i|-|s|}}, \tag{3}$$

where $MB(c_j)$ is the Markov Blanket of c_j. Here according to the Markov property, conditioned by $X_{G\setminus c_j}$ is equal to conditioned by $X_{MB(c_j)}$.

The scopes of the factors in MRF-F are in fact over all the variables regarding the default global configuration. The scopes of the factors in RMRF-F are $\{c_i \bigcup MB(c_i)\}$.

CRF-F (Lafferty *et al.*, 2001) can be considered as a special MRF-F, which factorizes the conditional probability. The chain structured CRF-F can be written in non-exponential form as follows:

$$P(y_1, y_2, ..., y_n | X) = \prod_{i=1}^{n-1} \phi_i(y_i, y_{i+1}, X) \prod_{i=1}^{n} f_i(y_i, X) \tag{4}$$

The transition feature functions $\{\phi_i\}$ are defined over edges $\{(y_i, y_{i+1})\}$ conditioned by X and the state feature functions $\{f_i\}$ are defined over nodes $\{y_i\}$ conditioned by X.

There are several questions arising naturally: (i) Conditional probability is used to factorize directed graph, then is there existing the equivalent for undirected graph? Intuitively, this equivalent should be symmetrical. (ii) What are the $f_i(y_i, X)$ and $\phi_i(y_i, y_{i+1}, X)$ in Eqn.(4) indeed? Could they be written as exact probability functions? In MRF, they are explained as "compatibility". But this vague intuition is far from a precise definition; (iii) Is there existing a unified mathematical foundation for all of these factorizations?

In this paper, we answer these questions by constructing a novel mathematical concept co-occurrence rate (CR). As CR-F is directly based on the independence relations among random variables, it can be directly applied to different types of PGMs, as different types of PGMs are just different representations of independence relations. CR has simple probability definition and clear intuitive interpretation. More important, we show that BN-F, CRF-F, MRF-F and RMRF-F are all special cases of CR-F. Thus CR provides a unified mathematical foundation for factorizing PGMs. CR-F can tell us not only the scopes of the factors, but also the exact probability functions of these factors. In CR-F, each factorizing step corresponds to a graph operation. CR-F can be visualized as applying a sequence of graph operations, including partition, merge, duplicate and condition, to the PGM graph. As "Graphical models are a marriage between probability theory and graph theory" (Jordan, 1998), the strong association between probability factorization and graph operations is a big advantage of CR-F. We also describe a systematic way to factorize TCG graphs into factors whose scopes are exactly over maximal cliques without any default configuration. This improves the results of (R)MRF-F and also indicates that (R)MRF-F, as special cases of CR-F, can not always achieve the optimal results of CR-F.

The remainder of paper is organized as follows: in Section (2), CR is developed. In Section (3), examples are given to demonstrate the CR-F for different types of PGMs. We also show that BN-F and CRF-F are special cases of CR-F. In Section (4), we show that (R)MRF-F are special cases of CR-F. Section (5) gives a systematic way to factorize TCGs. Conclusion, discussion and future work follow in the last two Sections (6, 7).

2 Development of CR

In this section, we construct the novel mathematical concept co-occurrence rate (CR) upon the foundations of probability theory. The concept of CR was inspired by Lenz-Ising model (Ising, 1925).

2.1 Definition of CR

CR between two events A and B is defined as:

$$CR(A, B) = \frac{P(A, B)}{P(A)P(B)},$$

where P is probability. CR can be intuitively interpreted as the interaction between the occurrences of A and B: (i) If $CR(A, B) = 1$, the occurrences of A and B are **independent**; (ii) If $CR(A, B) > 1$, the occurrences of A and B are **attractive**; (iii) If $0 \leq CR(A, B) < 1$, the occurrences of A and B are **repulsive**.

CR for discrete random variables is defined as:

$$CR(x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n)}{P(x_1)P(x_2)...P(x_n)}, \quad (5)$$

For the continuous random variables, we use the probability density function p:

$$\begin{aligned} &CR(x_1, x_2, ..., x_n) \\ &= \lim_{\varepsilon \downarrow 0} \frac{P(x_1 - \epsilon_1 \le x_1 \le x_1 + \epsilon_1, ..., x_n - \epsilon_n \le x_n \le x_n + \epsilon_n)}{P(x_1 - \epsilon_1 \le x_1 \le x_1 + \epsilon_1)...P(x_n - \epsilon_n \le x_n \le x_n + \epsilon_n)} \\ &= \lim_{\varepsilon \downarrow 0} \frac{\int_{x_1-\epsilon_1}^{x_1+\epsilon_1} ... \int_{x_n-\epsilon_n}^{x_n+\epsilon_n} p(x_1, ..., x_n) dx_1 ... dx_n}{\int_{x_1-\epsilon_1}^{x_1+\epsilon_1} p(x_1)dx_1 ... \int_{x_n-\epsilon_n}^{x_n+\epsilon_n} p(x_n)dx_n} \\ &= \lim_{\varepsilon \downarrow 0} \frac{2\epsilon_1 ... 2\epsilon_n p(x_1, ..., x_n)}{2\epsilon_1 p(x_1)...2\epsilon_n p(x_n)} = \frac{p(x_1, ..., x_n)}{p(x_1)...p(x_n)} \end{aligned}$$

In the rest of this paper, we only discuss the discrete situation. It can be easily extended to continuous random variables.

$$P(x_1, x_2, ..., x_n) = CR(x_1, x_2, ..., x_n)P(x_1)...P(x_n). \quad (6)$$

So instead of factorizing the joint probability, we can first factorize its CR, and then replace the CR in Eqn.(6) with the factorized CR. If there is only one random variable:

$$CR(x) = \frac{P(x)}{P(x)} = 1 \quad (7)$$

This can be intuitively explained as one event can happen independently by itself. But $CR(\emptyset) = \frac{P(\emptyset)}{P(\emptyset)}$ is undefined, as $P(\emptyset) = 0$.

Conditional probability can be written as CR functions:

$$\begin{aligned} P(x_1, x_2, ..., x_n|x) &= \frac{P(x_1, x_2, ..., x_n, x)}{P(x)} \quad (8) \\ &= CR(x_1, x_2, ..., x_n, x) \prod_{i=1}^{n} P(x_i). \end{aligned}$$

Notice that $CR(A, B, C) = \frac{P(A,B,C)}{P(A)P(B)P(C)}$ is **different** from $CR(A, BC) = \frac{P(A,BC)}{P(A)P(BC)} = \frac{P(A,B,C)}{P(A)P(BC)}$. The first CR means the co-occurrence rate among three events $\{A, B, C\}$. In the second one there are only two events : A and a joint event BC. But there is no such difference for P: $P(A, B, C) = P(A, BC) = P(ABC)$. This complies with the intuition of CR and P.

2.2 Definition of Conditional CR

The Conditional CR is defined as:

$$CR(x_1, ..., x_n|x) = \frac{P(x_1, ..., x_n|x)}{P(x_1|x)...P(x_n|x)}, \quad (9)$$

which is the co-occurrence rate of $\{x_1, ..., x_n\}$ conditioned by x. Then,

$$\begin{aligned} CR(x_1, ..., x_n|x) &= \frac{P(x_1, ..., x_n, x)P(x)^n}{P(x)P(x_1, x)...P(x_n, x)} \\ &= \frac{CR(x_1, ..., x_n, x)}{CR(x_1, x)...CR(x_n, x)}, \end{aligned}$$

and we can get the following theorem which allows the **condition operation** on the graph to deal with the incomplete graph as demonstrated in Section (3.4):

Condition Theorem

$$\begin{aligned} CR(x_1, ..., x_n) &= \frac{P(x_1, ..., x_n)}{P(x_1)...P(x_n)} = \frac{\sum_x P(x_1, ..., x_n, x)}{P(x_1)...P(x_n)} \quad (10) \\ &= \frac{\sum_x CR(x_1, ..., x_n, x)P(x_1)...P(x_n)P(x)}{P(x_1)...P(x_n)} \\ &= \sum_x CR(x_1, ..., x_n|x)CR(x_1, x)...CR(x_n, x)P(x). \end{aligned}$$

2.3 Commutative

If we consider CR as a operation on a set of variables, then CR is commutative:

$$\begin{aligned} CR(x_{a(1)}, x_{a(2)}, ..., x_{a(n)}) &= \frac{P(x_{a(1)}, x_{a(2)}, ..., x_{a(n)})}{P(x_{a(2)})P(x_{a(2)})...P(x_{a(n)})} \\ &= \frac{P(x_{b(1)}, x_{b(2)}, ..., x_{b(n)})}{P(x_{b(2)})P(x_{b(2)})...P(x_{b(n)})} \\ &= CR(x_{b(1)}, x_{b(2)}, ..., x_{b(n)}), \end{aligned}$$

where a and b are different permutations of $(1, 2, ..., n)$. This commutative law is important because it allows us to partition or merge the graph in any way.

2.4 Marginal CR

Random variables in CR can be eliminated by marginally summing up:

$$\begin{aligned} &\sum_{x_n} CR(x_1, x_2, ..., x_{n-1}, x_n)P(x_n) \\ &= \sum_{x_n} \frac{P(x_1, x_2, ..., x_{n-1}, x_n)}{P(x_1)P(x_2)...P(x_{n-1})P(x_n)} P(x_n) \\ &= \sum_{x_n} \frac{P(x_1, x_2, ..., x_{n-1}, x_n)}{P(x_1)P(x_2)...P(x_{n-1})} \\ &= \frac{P(x_1, x_2, ..., x_{n-1})}{P(x_1)P(x_2)...P(x_{n-1})} \\ &= CR(x_1, x_2, ..., x_{n-1}) \end{aligned}$$

If $n = 2$:

$$\sum_{x_2} CR(x_1, x_2)P(x_2) = CR(x_1) = 1.$$

where $CR(x_1) = 1$ by Eqn.(7).

2.5 Bi-partition Theorem

This is the critical theorem which allows the **bi-partition operation** on the graph to factorize a CR into three parts (the left, the right and the cut between the left and right):

$$\begin{aligned} &CR(x_1, ..., x_k, x_{k+1}, ..., x_n) \quad (11) \\ &= CR(x_1, ..., x_k)CR(x_{k+1}, ..., x_n)CR(x_1...x_k, x_{k+1}...x_n) \end{aligned}$$

This theorem can be proved as follows:

$$\begin{aligned}
&CR(x_1,...,x_k)CR(x_{k+1},...,x_n)CR(x_1...x_k,x_{k+1}...x_n)\\
&=\frac{P(x_1,...,x_k)}{\prod_{i=1}^k P(x_i)}\frac{P(x_{k+1},...,x_n)}{\prod_{i=k+1}^n P(x_i)}\\
&\ \ \frac{P(x_1,...,x_k,x_{k+1},...,x_n)}{P(x_1,...,x_k)P(x_{k+1},...,x_n)}\\
&=\frac{P(x_1,...,x_n)}{P(x_1)P(x_2)...P(x_n)}\\
&=CR(x_1,...,x_k,x_{k+1},...,x_n)
\end{aligned}$$

Bi-partition Theorem can be recursively used to further factorize the new CRs.

2.6 Merge Theorem

This theorem allows the **merge operation** which is inverse to partition operation.

$$\begin{aligned}
&CR(x_1,...,x_k,x_{k+1},...,x_n)\\
&=CR(x_1,...,x_kx_{k+1},...x_n)CR(x_k,x_{k+1})
\end{aligned} \tag{12}$$

where two subgraphs x_k and x_{k+1} are merged into one part x_kx_{k+1} and a new factor $CR(x_k,x_{k+1})$ is generated. This theorem can be proved as:

$$\begin{aligned}
&CR(x_1,...,x_kx_{k+1},...x_n)CR(x_k,x_{k+1})\\
&=\frac{P(x_1,...,x_n)}{P(x_1)...P(x_kx_{k+1})...P(x_n)}\frac{P(x_kx_{k+1})}{P(x_k)P(x_{k+1})}\\
&=\frac{P(x_1,...,x_kx_{k+1},...,x_n)}{P(x_1)...P(x_k)P(x_{k+1})...P(x_n)}\\
&=CR(x_1,...,x_k,x_{k+1},...,x_n)
\end{aligned}$$

There is a corollary following directly from this Merge Theorem and the Independence Theorem (Eqn.14):

if $(x_k \perp x_{k+1})$, then:

$$CR(x_1,...,x_k,x_{k+1},...,x_n)=CR(x_1,...,x_kx_{k+1},...x_n)$$

That is merging two independent random variables does not affect the global CR value.

2.7 Duplicate Theorem

This theorem allows **duplicate operation** to duplicate a random variable which already exists in the CR. This theorem is very useful when we manipulate overlapping subgraphs:

$$CR(x_1,..,x_i,..,x_n)=CR(x_1,..,x_i,x_i,..,x_n)P(x_i). \tag{13}$$

This theorem can be proved as follows:

$$\begin{aligned}
&CR(x_1,x_2,...,x_i,x_i,...,x_n)P(x_i)\\
&=\frac{P(x_1x_2...x_n)}{P(x_1)P(x_2)...P(x_i)P(x_i)...P(x_n)}P(x_i)\\
&=\frac{P(x_1x_2...x_n)}{P(x_1)P(x_2)...P(x_i)...P(x_n)}\\
&=CR(x_1,x_2,...,x_i,...,x_n).
\end{aligned}$$

2.8 Independence Theorem

If $\{x_1,x_2,...,x_n\}$ are mutually independent:

$$CR(x_1,x_2,...,x_n)=1. \tag{14}$$

2.9 Conditional Independence Theorems

2.9.1 The First CIT

If $(x_1x_2...x_k \perp y_1y_2...y_l|w_1w_2...w_m)$, then:

$$\begin{aligned}
&CR(x_1x_2...x_k,y_1y_2...y_lw_1w_2...w_m)\\
&=CR(x_1x_2...x_k,w_1w_2...w_m).
\end{aligned} \tag{15}$$

This theorem is used to reduce the random variables after a partition or merge operation. This theorem can be proved as:

$$\begin{aligned}
&(x_1x_2...x_k \perp y_1y_2...y_l|w_1w_2...w_m) \Rightarrow\\
&P(x_1...x_ky_1...y_lw_1...w_m)\\
&=\frac{P(x_1...x_kw_1...w_m)P(y_1...y_lw_1...w_m)}{P(w_1...w_m)}
\end{aligned}$$

then,

$$\begin{aligned}
&CR(x_1x_2...x_k,y_1y_2...y_lw_1w_2...w_m)\\
&=\frac{P(x_1...x_ky_1...y_lw_1...w_m)}{P(x_1...x_k)P(y_1...y_lw_1...w_m)}\\
&=\frac{P(x_1...x_kw_1...w_m)}{P(x_1...x_k)P(w_1...w_m)}\\
&=CR(x_1x_2...x_k,w_1w_2...w_m)
\end{aligned}$$

2.9.2 The Second CIT

If $(x_1x_2...x_k \perp y_1y_2...y_l|w_1w_2...w_m)$, then:

$$\begin{aligned}
&CR(w_1w_2...w_m,x_1x_2...x_ky_1y_2...y_l)\\
&=\frac{CR(x_1x_2...x_k,w_1w_2...w_m)CR(y_1y_2...y_l,w_1w_2...w_m)}{CR(x_1x_2...x_k,y_1y_2...y_l)}
\end{aligned}$$

This theorem is useful, because each CR on the right side has fewer random variables than the left CR.

$$\begin{aligned}
&CR(w_1w_2...w_m,x_1x_2...x_ky_1y_2...y_l)\\
&=\frac{P(w_1w_2...w_mx_1x_2...x_ky_1y_2...y_l)}{P(w_1w_2...w_m)P(x_1...x_ky_1...y_l)}\\
&=\frac{P(w_1...w_mx_1...x_k)P(w_1...w_my_1...y_l)}{P(w_1...w_m)P(w_1...w_m)P(x_1...x_ky_1...y_l)}\\
&=\frac{CR(x_1x_2...x_k,w_1w_2...w_m)CR(y_1y_2...y_l,w_1w_2...w_m)}{CR(x_1x_2...x_k,y_1y_2...y_l)}
\end{aligned}$$

2.9.3 The Third CIT

If $(x_1x_2...x_k \perp y_1y_2...y_l|w_1w_2...w_m)$, then:

$$\begin{aligned}
&CR(w_1...w_mx_1...x_k,w_1...w_my_1...y_l)\\
&=CR(w_1...w_m,w_1...w_m)\\
&=\frac{1}{P(w_1w_2...w_m)}
\end{aligned} \tag{16}$$

This theorem is useful when we deal with the overlapping clusters.

$$
\begin{aligned}
&CR(w_1...w_m x_1...x_k, w_1...w_m y_1...y_l) \\
&= \frac{P(w_1...w_m x_1...x_k y_1...y_l)}{P(w_1...w_m x_1...x_k)P(w_1...w_m y_1...y_l)} \\
&= \frac{1}{P(w_1...w_m)} = \frac{P(w_1...w_m)}{P(w_1...w_m)P(w_1...w_m)} \\
&= \frac{P(w_1...w_m, w_1...w_m)}{P(w_1...w_m)P(w_1...w_m)} \\
&= CR(w_1...w_m, w_1...w_m)
\end{aligned}
$$

2.10 Unconnected Nodes Theorem (UNT)

Suppose $\{a, b\}$ are two unconnected nodes in G. That is there is no direct edge between a and b. Then $a \perp b|MB(a, b)$, where $MB(a, b)$ is the Markov blanket of $\{a, b\}$. And suppose $W, X \in \mathcal{P}(G\backslash\{a, b\})$ including the boundary cases $\{\emptyset, G\backslash\{a, b\}\}$, $MB(a, b) \subseteq W \cup X$, and $W \cap X = \emptyset$. Then $(a \perp b|W, X)$ and we get the UNT:

$$
\begin{aligned}
&CR(W, a = 0, b = 0, X = 0)CR(W, a, b, X = 0) \quad (17)\\
&= CR(W, a = 0, b, X = 0)CR(W, a, b = 0, X = 0)
\end{aligned}
$$

For the left side, we partition (Eqn.11) a out and apply the first CIT (Eqn.15):

$$
\begin{aligned}
&CR(W, a = 0, b = 0, X = 0)CR(W, a, b, X = 0) \\
&= CR(W, b = 0, X = 0)CR(a = 0, WX = 0) \\
&\quad CR(W, b, X = 0)CR(a, WX = 0)
\end{aligned}
$$

For the right side, we also partition a out and apply the first CIT:

$$
\begin{aligned}
&CR(W, a = 0, b, X = 0)CR(W, a, b = 0, X = 0) \\
&= CR(W, b, X = 0)CR(a = 0, WX = 0) \\
&\quad CR(W, b = 0, X = 0)CR(a, WX = 0)
\end{aligned}
$$

As the left side equals the right side, we proved the theorem.

3 Examples

In this section, we demonstrate CR-F on different PGMs based on the results obtained in Section (2).

3.1 Example 1: A Bayesian Network

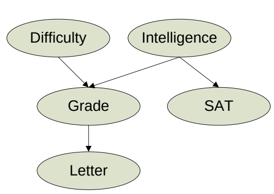

Figure 1: A BN (Koller & Friedman, 2009)

Fig.(1) is a Bayesian network. By Eqn.(6):

$$
\begin{aligned}
&P(D, I, G, S, L) \\
&= CR(D, I, G, S, L)P(D)P(I)P(G)P(S)P(L).
\end{aligned}
$$

We go on to factorize $CR(D, I, G, S, L)$. Factorization using CR is to apply a sequence of graph operations including partition, merge, duplicate and condition to the graph. After each operation, we check if the CITs in Section (2.9) can be applied to reduce random variables. As there are a lot of such operation sequences, consequently we can get a lot of different factorization results. All of them are mathematically correct[1]. We illustrate two of them as follows:

Factorization 1 (by partition):

Step1: $(\{D, I, G, S, L\}) \rightarrow (\{D\}, \{I, G, S, L\})$.

$$
\begin{aligned}
CR(D, I, G, S, L) &= CR(D)CR(I, G, S, L)CR(D, IGSL) \\
&= CR(I, G, S, L)CR(D, G)
\end{aligned}
$$

We get the first equation by partition operation (Eqn.11). We get the second equation by the First CIT (Eqn.15) as $(D \perp ISL|G)$. And $CR(D) = 1$ (Eqn.7).

Step2: $(\{I, G, S, L\}) \rightarrow (\{S\}, \{I, G, L\})$.

$$CR(I, G, S, L) = CR(I, G, L)CR(S, I)$$

Step3: $(\{I, G, L\}) \rightarrow (\{I\}, \{G, L\})$.

$$CR(I, G, L) = CR(I, G)CR(G, L)$$

Finally:

$$CR(D, I, G, S, L) = CR(D, G)CR(S, I)CR(I, G)CR(G, L)$$

Factorization 1 (by merge):

Step1: $\{D, I, G, S, L\} \rightarrow \{D, I, S, GL\}$.

$$CR(D, I, G, S, L) = CR(D, I, S, GL)CR(G, L)$$

We get this equation by merge operation (Eqn.12).

Step2: $\{D, I, S, GL\} \rightarrow \{D, S, IGL\}$.

$$
\begin{aligned}
CR(D, I, S, GL) &= CR(D, S, IGL)CR(I, GL) \\
&= CR(D, S, IGL)CR(I, G)
\end{aligned}
$$

Step3: $\{D, S, IGL\} \rightarrow \{S, DIGL\}$.

$$
\begin{aligned}
CR(D, S, IGL) &= CR(S, DIGL)CR(D, IGL) \\
&= CR(S, I)CR(D, G)
\end{aligned}
$$

Finally:

$$CR(D, I, G, S, L) = CR(G, L)CR(I, G)CR(S, I)CR(D, G)$$

Factorization 2:

In the remainder of the paper, we only demonstrate factorization by partition. Factorization by merge can be easily obtained by merging the nodes in the reverse direction of factorization by partition.

[1] The logical consideration of the relation between BN-F and CR-F will be discussed in another paper.

Step1: $(\{D, I, G, S, L\}) \rightarrow (\{S\}, \{I, G, D, L\})$.

$$CR(D, I, G, S, L) = CR(I, G, D, L)CR(S, I).$$

Step2: $(\{I, G, D, L\}) \rightarrow (\{D, I\}, \{G, L\})$.

$$\begin{aligned} CR(I, G, D, L) &= CR(D, I)CR(G, L)CR(DI, GL) \\ &= CR(G, L)CR(DI, G) \end{aligned}$$

We get the second equation as $(D \perp I)$ (Eqn.14) and $(DI \perp L|G)$.

Finally:

$$CR(D, I, G, S, L) = CR(S, I)CR(DI, G)CR(G, L)$$

If we group the CRs in the above equation into proper scopes, we can get the result of BN-F:

$$\begin{aligned} &P(D, I, G, S, L) \\ &= CR(S, I)CR(DI, G)CR(G, L)P(D)P(I)P(G)P(S)P(L) \\ &= P(D)P(I)CR(DI, G)P(G)CR(S, I)P(S)CR(G, L)P(L) \\ &= P(D)P(I)P(G|DI)P(S|I)P(L|G) \end{aligned}$$

The factors in BN-F can be obtained by keeping all the fathers of a node in the same part when we are partitioning the graph. So BN-F can be considered as a special case of CR-F.

3.2 Example 2: Tree-Structured Markov Network

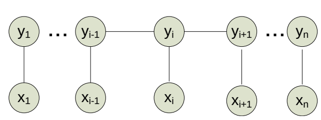

Figure 2: A Tree-Structured Markov Network

The tree-structured Markov network can be factorized by partitioning one leaf out each time. This results in the factors over all the edges and nodes.

$$\begin{aligned} &P(y_1, y_2, ..., y_n, x_1, x_2, ..., x_n) \\ &= CR(y_1, y_2, ..., y_n, x_1, x_2, ..., x_n) \prod_{i=1}^{n} P(y_i) \prod_{i=1}^{n} P(x_i) \\ &= \prod_{i=1}^{n} CR(x_i, y_i) \prod_{i=2}^{n} CR(y_{i-1}, y_i) \prod_{i=1}^{n} P(y_i) \prod_{i=1}^{n} P(x_i) \\ &= \prod_{i=1}^{n} \frac{P(x_i, y_i)}{P(x_i)P(y_i)} \prod_{i=2}^{n} \frac{P(y_{i-1}, y_i)}{P(y_{i-1})P(y_i)} \prod_{i=1}^{n} P(y_i) \prod_{i=1}^{n} P(x_i) \end{aligned}$$

3.3 Example 3: Chain-Structured CRF

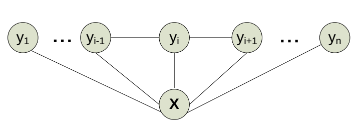

Figure 3: Chain-Structured CRF (Lafferty *et al.*, 2001)

CRF can be considered as a special MRF which is to factorize the conditional probability. Here we show that CRF-F is a special case of CR-F:

$$P(y_1, y_2, ..., y_n|X) = CR(y_1, y_2, ..., y_n|X) \prod_{i=1}^{n} P(y_i|X) \tag{18}$$

$$= \prod_{i=1}^{n-1} CR(y_i, y_{i+1}|X) \prod_{i=1}^{n} P(y_i|X) \tag{19}$$

$$= \prod_{i=1}^{n-1} \frac{P(y_i, y_{i+1}|X)}{P(y_i|X)P(y_{i+1}|X)} \prod_{i=1}^{n} P(y_i|X) \tag{20}$$

We get Eqn.(18) by Eqn.(9). We obtain Eqn.(19) from Eqn.(18) because under the condition X, $\{y_1, ..., y_n\}$ are chain structured and can be partitioned as Example 2. We can see that $CR(y_i, y_{i+1}|X)$ and $P(y_i|X)$ are just the transition feature functions and state feature functions in CRF-F (Eqn.4), respectively. CR-F tells us not only the scopes of the factors, but also the exact probability functions of these factors, where $\phi_i(y_i, y_{i+1}, X) = CR(y_i, y_{i+1}|X) = \frac{P(y_i, y_{i+1}|X)}{P(y_i|X)P(y_{i+1}|X)}$ and $f_i(y_i, X) = P(y_i|X)$. CRF-F can not tell us the exact probability functions of the factors.

3.4 Example 4: Arbitrary Markov Network

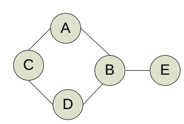

Figure 4: A Markov Network

In this example, we show how to factorize an arbitrary Markov network. Especially, we demonstrate how to deal with the incomplete graph by using the condition operation (Eqn.10):

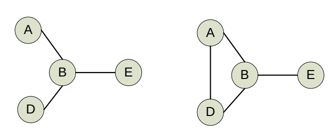

Figure 5: The Incomplete Structures

Factorization:

Step1: $(\{A, B, C, D, E\}) \rightarrow (\{C\}, \{A, B, D, E\})$.

$$CR(A, B, C, D, E) = CR(C, AD)CR(A, B, D, E)$$

Now come the incomplete structures $\{A, B, D, E\}$ as shown in Fig.(5). Should we go on to factorize $CR(A, B, D, E)$ using the left structure or the right structure? According to the independence semantics of the original graph, we have $(C \perp B|AD)$, $(A \perp D|BC)$ and $(E \perp AD|B)$. We have already used

the $(C \perp B|AD)$ at the first step. $(E \perp AD|B)$ is not related to C, so no matter the left structure or the right structure, it always holds. There are two choices for the $(A \perp D|BC)$. The left structure means under the condition C, $(A \perp D|B)$; and the right structure means $(A \not\perp D|B)$. Both of them are correct.

Step2(Left/Right): $\{A, B, D, E\} \to (\{A, B, D\}, \{E\})$.

$$CR(A, B, D, E) = CR(A, B, D)CR(D, E)$$

Step3(Left): $\{A, B, D|C\} \to (\{A, B|C\}, \{D|C\})$.

By the condition operation (Eqn.10):

$$\begin{aligned}
&CR(A, B, D) \\
&= \sum_C [CR(A, B, D|C)CR(A, C)CR(B, C)CR(D, C)P(C)] \\
&= \sum_C [CR(A, B|C)CR(B, D|C)CR(A, C)CR(B, C) \\
&\quad CR(D, C)P(C)]
\end{aligned}$$

Step3(Right): $\{A, B, D\} \to \{A, B, D\}$.

$$CR(A, B, D) = CR(A, B, D)$$

The results of Step3(Left) and Step3(Right) are equal regarding the independence semantics of the original graph in Fig.(4). With the condition operation we can utilize all conditional independences. In this example, if we did not use condition operation, then the conditional independence $(A \perp D|BC)$ could not be used.

4 CR-F and (R)MRF-F

Using CR-F, there can be a lot of different ways to factorize a graph. In this section, we show that the factors of (R)MRF-F can be obtained by a very special operation sequence of CR. Thus (R)MRF-F are just special cases of CR-F.

Suppose the nodes in G: $G = \{g_1, g_2, ..., g_n\}$. For each $S \in \mathcal{P}(G)\backslash G$ including $\emptyset$ repeat the following two steps for $2^{|G|-|S|-1}$ times:

1. Duplicate (Eqn.13) the nodes in G:

$$CR(G) = CR(G, G)P(g_1)...P(g_n)$$

2. Partition the G out:

$$\begin{aligned}
CR(G) &= CR(G, G)P(g_1)...P(g_n) \\
&= CR(G)CR(G, G)CR(G)P(g_1)...P(g_n) \\
&= CR(G)\frac{P(g_1)...P(g_n)}{P(g_1, ..., g_n)}CR(G) \\
&= \frac{CR(G)}{CR(G)}CR(G)
\end{aligned}$$

As $\frac{CR(G)}{CR(G)} = 1$, we can assign arbitrary values to $G\backslash S$, and we get:

$$CR(G) = \frac{CR(S, G\backslash S = 0)}{CR(S, G\backslash S = 0)}CR(G)$$

Then factorize the $CR(G)$ on the right side for the next S. And finally we get:

$$CR(G) = [\prod_{S \in \{\mathcal{P}(G)-G\}} (\frac{CR(S, G\backslash S = 0)}{CR(S, G\backslash S = 0)})^{2^{|G|-|S|-1}}]CR(G) \tag{21}$$

This equation seems pretty special (stupid?). Now in fact we have already obtained the factors in MRF-F by CR-F. What remained is to group these factors into proper scopes. The scopes are just all the subset of G: $\mathcal{P}(G)$ including $\emptyset$ and G. For each scope $S \in \mathcal{P}(G)$, we select the following factors in Eqn.(21) into S:

$$\{CR(W, G\backslash W = 0)^{(-1)^{|S|-|W|}}, W \in \mathcal{P}(S)\}.$$

We call these factors as W factors. The following two binomial equations guarantee that all the factors are just be selected into scopes $\mathcal{P}(G)$ in this way:

$$\begin{aligned}
&2^{|G|-|W|} = (1+1)^{|G|-|W|} \\
&= C^0_{|G|-|W|} + C^1_{|G|-|W|} + ... + C^{|G|-|W|}_{|G|-|W|}
\end{aligned} \tag{22}$$

$$\begin{aligned}
&0^{|G|-|W|} = (1-1)^{|G|-|W|} \\
&= C^0_{|G|-|W|} - C^1_{|G|-|W|} + ... + (-1)^{|G|-|W|}C^{|G|-|W|}_{|G|-|W|}
\end{aligned} \tag{23}$$

The number of W factors in Eqn.(21) is $2 * 2^{|G|-|W|-1} = 2^{|G|-|W|}$. Half of them are in numerator and the other half in denominator. W factors are included once by each of $\{S, W \subseteq S\}$. Eqn.(22) tells the number of $\{S\}$ which contain the W factor is also $2^{|G|-|W|}$, so all the W factors are just included into $\{S\}$. Eqn.(23) tells half of $\{S\}$ select the W factors in the numerator and the other half select the W factors in the denominator.

We go on to prove that if a scope S is not a clique, all the factors selected into S cancel themselves out:

1. If S is not a clique, then there must be two unconnected nodes $\{a, b\}$ in S.

2. Suppose $W \in \mathcal{P}(S\backslash\{a, b\})$. Thus all the subsets in S can be categorized into four types: W, $W \cup \{a\}$, $W \cup \{b\}$ and $W \cup \{a, b\}$. And they must be in the following form in the scope S:

$$\phi(S) = \prod_W [\frac{CR(W, a=0, b=0, X=0)CR(W, a, b, X=0)}{CR(W, a=0, b, X=0)CR(W, a, b=0, X=0)}]^{-1^*} \tag{24}$$

where $X = G\backslash\{W, a, b\}$. The absolute positions of these four factors are not important. We only need their relative positions are correct as they will cancel themselves out. So we denote the power as -1^*. As $MB(a, b) \subseteq W \cup X = G\backslash\{a, b\}$ and $W \cap X = \emptyset$, according to the UNT (Eqn.17), if we assign all the default values from an arbitrary but fixed global configuration, then $\phi(S) = 1$. □

Now only the factors in cliques are left. Cliques $\{c_i\}$ can be categorized into three types: $\emptyset$, $|c_i| = 1$ and

$|c_i| \geq 2$. The factor in the empty clique is: $CR(G=0)$; the factors in one node clique are: $\frac{CR(g_i, G\backslash g_i=0)}{CR(g_i=0, G\backslash g_i=0)}$, where g_i is the unique node in this clique; and factors in multi-node cliques can be written in the same form as Eqn.(24), where $\{a,b\}$ can be any pair of nodes in the clique. Then

$$
\begin{aligned}
&P(g_1,...,g_n) = CR(g_1,...,g_n)P(g_1)...P(g_n) \\
&= CR(G=0)\prod_{i=1}^{n}\frac{CR(g_i, G\backslash g_i=0)P(g_i)}{CR(g_i=0, G\backslash g_i=0)} \qquad (25) \\
&\prod_{|c_i|\geq 2}\prod_{w}[\frac{CR(w,a=0,b=0,X=0)CR(w,a,b,X=0)}{CR(w,a=0,b,X=0)CR(w,a,b=0,X=0)}]^{-1^{|c_i|-|w|}},
\end{aligned}
$$

where $w \in \mathcal{P}(c_i\backslash\{a,b\})$ and $X = G\backslash\{w,a,b\}$. If we substitute the CRs in Eqn.(25) with their probability definition (Eqn.5), we get MRF-F (Eqn.1) exactly. As we can obtain the factors in MRF-F by CR-F, MRF-F can be considered as a special case of CR-F.

We can further refine the scopes in Eqn.(25):

$$
\begin{aligned}
&\frac{CR(g_i, G\backslash g_i=0)}{CR(g_i=0, G\backslash g_i=0)} = \frac{CR(g_i, MB(g_i)=0, X=0)}{CR(g_i=0, MB(g_i)=0, X=0)} \\
&= \frac{CR(g_i)CR(g_i, MB(g_i)=0)CR(MB(g_i)=0, X=0)}{CR(g_i=0)CR(g_i=0, MB(g_i)=0)CR(MB(g_i)=0, X=0)} \\
&= \frac{CR(g_i, MB(g_i)=0)}{CR(g_i=0, MB(g_i)=0)}, \qquad (26)
\end{aligned}
$$

where $X = G\backslash\{g_i, MB(g_i)\}$. And also:

$$
\begin{aligned}
&\frac{CR(w,a=0,b=0,X=0)CR(w,a,b,X=0)}{CR(w,a=0,b,X=0)CR(w,a,b=0,X=0)} \\
&= \frac{CR(w,a=0,b=0,M=0,N=0,H=0)}{CR(w,a=0,b,M=0,N=0,H=0)} \\
&\quad \frac{CR(w,a,b,M=0,N=0,H=0)}{CR(w,a,b=0,M=0,N=0,H=0)} \\
&= \frac{CR(w,a=0,b=0,M=0,N=0)}{CR(w,a=0,b,M=0,N=0)} \\
&\quad \frac{CR(w,a,b,M=0,N=0)}{CR(w,a,b=0,M=0,N=0)} \qquad (27)
\end{aligned}
$$

where $M = c\backslash\{w,a,b\}$, $N = MB(c)$ and $H = G\backslash\{c \cup MB(c)\}$. If we first replace Eqn.(25) with Eqn.(26) and Eqn.(27), and then replace the CRs in the new equation using Eqn.(8) with $N=0$ as the condition, we get RMRF-F(Eqn.3) exactly. We can see that in the refinement steps Eqn.(26) and Eqn.(27), we just further applied the partition operations and first CIT to the existing factors. That means we can get the factors of RMRF-F by a sequence of graph operations. Therefore RMRF-F is a special case of CR-F.

5 Factorizing TCG

In this section, we describe a systematic way to factorize TCGs into factors which are defined exactly over the maximal cliques without any default configuration. First, we review the concept of clique graph (Hamelink, 1968) in graph theory.

*The **clique graph (CG)** of a given graph $G(V,E)$ is a graph $G'(V',E')$. The nodes of G' are defined as $V' = \{C_1, C_2, ..., C_n\}$. There exists a one-to-one mapping between $\{C_1, C_2, ..., C_n\}$ and all the maximal cliques $\{c_1, c_2, ..., c_n\}$ in G. The edges in G' are defined as $E' = \{(C_i, C_j); V(c_i) \cap V(c_j) \neq \emptyset; 1 \leq i,j \leq n; i \neq j\}$.*

Here we define Tree structured CG (TCG) by Alg.(1). Notice that according to our definition whether a CG is TCG can not be simply judged by existence of cycles in the CG. Even a CG contains cycles, it may also be a TCG as the example shown in Fig.(6).

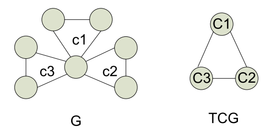

Figure 6: A graph and its TCG

Algorithm 1 isTCG

Input: $G(V,E)$ and its CG $G'(V',E')$
while *true* **do**
  if $|V'| \leq 1$ **then**
    return *true*;
  end if
  $noChange = true$;
  for $i = 1$ **to** $|V'|$ **do**
    {Here $adj(C_i) = \{C_k, ..., C_l\}$ are all the adjacent nodes of C_i in G' and $\{c_k, ..., c_l\}$ are their corresponding maximal cliques in G.}
    if $\exists C_j \; \forall C_h \; c_i \cap c_h \subseteq c_i \cap c_j; C_j, C_h \in adj(C_i)$ **then**
      $G' = G' - C_i$;
      $noChange = false$;
      break;
    end if
  end for
  if $noChange == true$ **then**
    return $false$;
  end if
end while

TCGs can be factorized as follows:

Step0: $P(x_1, ..., x_{|V|}) = CR(x_1, ..., x_{|V|})P(x_1)...P(x_{|V|})$.

Step1: Select a node C_i, for which $\exists C_j \; \forall C_h \; c_i \cap c_h \subseteq c_i \cap c_j; C_j, C_h \in adj(C_i)$. We call C_j as maximum adjacent node of C_i and denoted as $Maxadj(C_i)$. Alg.(1) guarantees that for a TCG there always exists such a node during the factorization process. Duplicate $\{x_k, ..., x_l\} = V(c_i) \cap V(c_j)$:

$$CR(x_1, ..., x_{|V|}) = CR(x_1, ..., x_{|V|}, x_k, ..., x_l)P(x_k)...P(x_l)$$

Step2: Then we partition the random variables $\{x_1, x_2, ..., x_{|v|}, x_k, ..., x_l\}$ into two parts: $\{x_p, ..., x_q\} =$

$V(c_i)$ and the remainder $\{x_h, ..., x_m\} = \cup V(c \backslash c_i)$.

$$
\begin{aligned}
&CR(x_1, x_2, ..., x_{|v|}, x_k, ..., x_l) \\
&= CR(x_p, ..., x_q) CR(x_h, ..., x_m) CR(x_p...x_q, x_h...x_m) \\
&= CR(x_p, ..., x_q) CR(x_h, ..., x_m) CR(x_k...x_l, x_k...x_l) \quad (28) \\
&= CR(x_p, ..., x_q) CR(x_h, ..., x_m) \frac{1}{P(x_k, ..., x_l)}. \quad (29)
\end{aligned}
$$

We obtain Eqn.(28) from Eqn.(16). $\{x_k...x_l\}$ completely separate c_i from the remainder of G, so $(x_p...x_{k-1}x_{l+1}...x_q \perp x_h...x_{k-1}x_{l+1}...x_m | x_k...x_l)$. As in Eqn.(29) $\{x_p, ..., x_q\} = V(c_i)$ and $\{x_k, ..., x_l\} = V(c_i) \cap V(c_j)$, the scope of $\frac{CR(x_p, ..., x_q)}{P(x_k...x_l)}$ is just the maximal clique c_i. Repeat Step1 and Step2 until only one clique left:

$$
\begin{aligned}
&P(x_1, x_2, ..., x_{|v|}) \\
&= \prod_{i=1}^{|V'|-1} [\frac{CR(V(c_i))}{P(V(c_i) \cap V(Maxadj(c_i)))} \prod_{x_i \in V(c_i)} P(x_i)] \\
&\quad CR(V(c_{|V'|})) \prod_{x_i \in V(c_{|V'|})} P(x_i),
\end{aligned}
$$

where $C_{|V'|}$ is the root of G', which is the final clique left in Alg.(1). Therefore the probability functions over maximal cliques can be written as follows:
If C_i is not the root of G':

$$
\begin{aligned}
\phi_i(c_i) = &\frac{CR(V(c_i))}{P(V(c_i) \cap V(Maxadj(c_i)))} \prod_{x_i \in V(c_i)} P(x_i) \\
= &\frac{P(V(c_i))}{P(V(c_i) \cap V(Maxadj(c_i)))}.
\end{aligned}
$$

If C_i is the root of G':

$$
\phi_i(c_i) = CR(V(c_i)) \prod_{x_i \in V(c_i)} P(x_i) = P(V(c_i)).
$$

6 Conclusion

In this paper, we constructed the novel mathematical concept CR upon the foundations of probability theory. CR provides a unified mathematical foundation for factorizing PGMs. We illustrated that BN-F, CRF-F, MRF-F and RMRF-F are all special cases of CR-F. The factors of CR-F can be written as exact probability functions. We described a systematic way to factorize TCG with factor scopes exactly over maximal cliques without any default configuration, which improves the results of (R)MRF-F.

7 Discussion and Future Work

In this paper, we focussed on constructing the mathematical foundation for factorizing PGMs and do not mention learning and inference methods. But please notice that as BN-F, CRF-F, MRF-F and RMRF-F are all special cases of CR-F, the learning and inference methods based on the results of these factorizations can also be applied to CR-F. Using CR-F, we may get factorizations that consist of much fewer factors defined on local scopes. And more important, these factors can be written as exact probability functions. This should benefit learning and inference.

References

Abbeel, P., Koller, D., & Ng, A. (2005). Learning factor graphs in polynomial time & sample complexity. In *UAI-05*. Arlington, Virginia: AUAI Press, 1–9.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Statistical Science, 1 ed.

Cheung, S. (2008). Proof of Hammersley-Clifford Theorem. Tech. rep.

Clifford, P. (1990). Markov random fields in statistics. In *Disorder in Physical Systems*. 19–32.

Hamelink, R. C. (1968). A partial characterization of clique graphs. In *Journal of Combinational Theory*. 192–197.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. In *Zeitschrift für Physik A Hadrons and Nuclei*. vol. 31, 253–258.

Jordan, M. I. (1998). *Learning in Graphical Models*. MIT Press.

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML-01*. 282–289.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. In *Artificial Intelligence*. vol. 29, 241–288.